Cyberbullying Detection using Time Series Modeling

Nektaria Potha
Department of Information and Communication Systems Engineering
University of the Aegean
Samos, Greece
nekpotha@aegean.gr

Manolis Maragoudakis
Department of Information and Communication Systems Engineering
University of the Aegean
Samos, Greece
mmarag@aegean.gr

***Abstract*— Cyberbullying is a new phenomenon resulting from the advance of new communication technologies including the Internet, cell phones and Personal Digital Assistants. It is a challenging bullying problem occurring in a new territory. Online bullying can be particularly damaging and upsetting because it's usually anonymous or hard to trace. In this paper, the proposed method is utilizing a dataset of real world conversations (i.e. pairs of questions and answers between cyber predator and the victim), in which each predator question is manually annotated in terms of severity using a numeric label. We approach the issue as a sequential data modelling approach, in which the predator's questions are formulated using a Singular Value Decomposition representation. The motivation of this procedure is to study the accuracy of predicting the level of cyberbullying attack using classification methods and also to examine potential patterns between the lingustic style of each predator. More specifically, unlike previous approaches that consider a fixed window of a cyber-predator's questions within a dialogue, we exploit the whole question set and model it as a signal, whose magnitude depends on the degree of bullying content. Using feature weighting and dimensionality reduction techniques, each signal is straightforwardly parsed by a neural network that forecasts the level of insult within a question given a window between two and three previous questions. Throughout the time series modeling experiments, an interesting discovery was made. By applying SVD on the time series data and taking into account the second dimension (since the first is usually modeling trivial dependencies between instances and attributes) we observed that its plot was very similar to the plot of the class attribute. By applying a Dynamic Time Warping algorithm, the similarity of the aforementioned signals was proved to exist, providing an immediate indicator for the severity of cyberbullying within a given dialogue.**

Keywords—cyberbullying; time series analysis; singular value decomposition; SVM feature selection; dynamic time warping

I. INTRODUCTION

Cyberbullying is bullying that takes place using electronic technology. It is defined as a person tormenting, threatening, harassing or embarrassing another person using the internet or other technologies, like cell phones. Examples of cyberbullying include mean text messages or emails, rumors sent by email or posted on social networking sites, as well as embarrassing pictures, videos, websites, or fake profiles. In other words, cyberbullying is anything that gets posted online and is deliberately intended to hurt. In its most basic sense bullying involves two people, a bully or intimidator or predator and a victim. One main factor that have intensified this social menace is the anonymity of technology which allows cyber-predators to constantly reach their targets (victims) to do so more intensely than regular bullying with a lessened sense of responsibility [1].

This relatively new phenomenon of cyberbullying has proven to be difficult for researchers to study due to its "intangible, no corporeal nature. Ironically, it is this intangible, non-corporeal nature that makes electronic mediums ideal for bullies: there is an ease with which one may disguise themselves as another or remain entirely anonymous while online, with the ability to "hide" behind the computer screen [2],[3]. Temporary e-mail accounts and pseudonyms in chat rooms, instant messaging programs, and other Internet venues can make it very difficult for adolescents to determine the identity of aggressors (e.g. an unknown nickname is often used).

In the present paper, data mining methodologies are applied to the issue of cyberbullying detection. Utilizing a dataset of real world dialogues (pairs of questions and answers between cyber-predator and the victim) in which each question is manually annotated in terms of severity, we model each set of predator's questions as a time series. This is the first novelty of the proposed study, since the majority of researchers use either a predefined window of previous predator's questions, or do not give emphasis on the dialogue course and consider each question as an independent individual. An interesting aspect of this dataset is the severity is not only captured by sings of swearing or offending words but also from behavioral patterns. Section IV describes the characteristics of the dataset in more details.

The figure below (Fig. 1) portrays the architecture of the proposed method. The feature space is represented as a bag-of-words, experimenting with various representation measurements such as *tf-idf*, term frequency, term occurrence and binary occurrence. Given the large amount of extracted features, we applied feature selection by weighting each term using a Support Vector Machines (SVM) approach. Results on the forecasting performance of a neural net, supported out initial expectations that SVM are well suited for reducing the number of feature by weighting their importance to the class attribute. Experiments were also carried out using feature reduction techniques such as Singular Value Decomposition (SVD). Again, the forecasting performance was significantly higher

978-1-4799-4274-9/14 $31.00 © 2014 IEEE
DOI 10.1109/ICDMW.2014.170

than that of standard bag-of-words. However, modeling the time series using SVD led to the second novelty on the article, i.e. the discovery of similar behavior of the class signal and the signal of the second SVD dimension. Since the first SVD dimension measures the similarity of instances in terms of document length and the similarity of attributes in terms of frequency, it is expected that such trivial patterns usually do not reveal any unknown and important patterns within data.

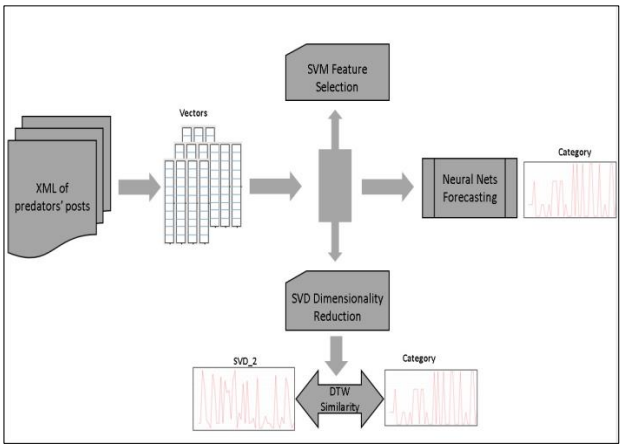

Fig. 1. The methodology used in the current study as a flowchart.

Furthermore, when plotting the second SVD dimension against the class attribute, the two signals portrayed similar behavior, which was confirmed when applied a Dynamic Time Warping (DTW) algorithm. The identification of such a correlation between the aforementioned signals could assist the dialogue annotation process or even be used to identify repeated offenders that use the same dialogue style in their attacks.

The rest of this paper is organized as follows: Section 2 presents previous work in cyberbullying detection while Section 3 provides some theoretical background for the data mining techniques used within the current research. Section 4 describes the process of modeling dialogues as time series and analyzing signals using feature reduction and forecasting approaches. In Section 5 discusses the experimental evaluation phase using the dataset described earlier while Section 5 includes the main conclusions drawn from this study and discusses future work directions.

II. Previous Work

As mentioned earlier, Cyberbullying is a growing problem in the social web and is becoming a major threat to teenagers and adolescents. The textual contents in a web environment is often unstructured, informal, and even misspelled. Moreover, the fact that predators attempt to use a language style that mimics that of teenagers, therefore vocabularies and standard text processing would perform poorly, makes the detection of cyberbullying a very hard process. As a result, few research teams are working on this task.

A Sexual Predator Identification competition took place for the first time at PAN-2012. Given a set of chat logs the participants had to identify the predators among all users in the different conversations or the part (the lines) of the conversations which are the most distinctive of the predator behavior. In conclusion, it is impossible to identify predators using a unique method but it is necessary the use of different approaches. Moreover the most effective method for identifying distinctive lines of the predator behavior in a chat log appeared to be those based on filtering on a dictionary or LM basis [4].

Yin et al., was the sole submission in the misbehavior detection task of CAW 2.0 [5]. Using three from the five datasets which were provided by the organizers of the content analysis workshop, they proposed a supervised learning approach for detecting harassment with a focus on detecting intentional annoyance. By employing a SVM classifier with the linear kernel and combining *tf-idf* measure as local features, sentiment features, and contextual features of documents proved that identification of online harassment provide significantly improved performance when *tf-idf* is supplemented with sentiment and contextual feature attributes. The results show improvements over the baselines.

In a recent study on cyberbullying detection, Kontostathis et al., [6] taking a collection of posts from the website Formspring.me, which allows users to post questions anonymously (a question- answer website where users openly invite others to ask and answer questions) proposed a "bag-of-words" language model, which based on the text in online posts, in order to detect instances of cyberbullying. Moreover, they exploited a supervised machine learning called Essential Dimensions of LSI (EDLSI) approach in order to identify additional terms of cyberbullying in Formspring.me data.

The data was labeled using a web service, Amazon's Mechanical Turk. The Mechanical Turk (MTurk) is a crowdsourcing Internet marketplace that enables individuals or businesses (known as *Requesters*) to co-ordinate the use of human intelligence to perform tasks that computers are currently unable to do. It is one of the sites of Amazon Wed Services. The goal was to identify the most commonly used cyberbullying terms. The Requesters are able to post tasks known as HITs (Human Intelligence Tasks), such as choosing the best among several photographs of a store-front, writing product descriptions, or identifying performers on music CDs. *Workers* (called *Providers* in Mechanical Turk's Terms of Service, or, more colloquially, *Turkers*) can then browse among existing tasks and complete them for a monetary payment set by the Requester. Requesters, (which are typically businesses) pay 10 percent of the price of successfully completed HITs to Amazon [7].

Latest cyber-security studies show that there is a rapid growth in the number of cybercrimes which cause tremendous financial losses to click-and-mortar organizations in recent years. The main contribution of the research work reported in this paper is the design of a novel, weakly supervised cybercriminal network mining method that is underpinned by a context-sensitive text mining enhanced probabilistic generative model. The proposed method is underpinned by a probabilistic generative model enhanced by a novel context-sensitive Gibbs

sampling algorithm. Evaluated based on two social media corpora, our experimental results reveal that the proposed method significantly outperforms the Latent Dirichlet Allocation (LDA) based method and the Support Vector Machine (SVM) based method by 5.23% and 16.62% in terms of Area Under the ROC Curve (AUC), respectively [8].

Cyberbullying detection tasks have mainly focused on the content of the conversations (of the text written by the participants, both the victim and the bully), rather than the features and characteristics of those involved. Dadvar et al., [9] proposed cyberbullying detection based on gender information. Using a supervised learning approach (SVM) in order to detect cyberbullying, they proved that taking gender-specific language features into account and categorizing into male and female groups improves the discrimination capacity of a classifier to detect cyberbullying.

The creation of a program which will be able to detect the occurrence of sexual predation in an online social setting was discussed by McGhee et al., The logic behind this approach is the existence of a rule-based system for labeling predatory posts in a chat transcript [10].

In general, the majority of popular social media use simple lexicon-based approach to filter offensive contents. Their lexicons are either predefined (such as Youtube) or composed by the users themselves (such as Facebook). Furthermore, most sites rely on users to report offensive contents to take actions. Because of their use of simple lexicon-based automatic filtering approach to block the offensive words and sentences, these systems have low accuracy and may generate many false positive alerts. In addition, when these systems depend on users and administrators to detect and report offensive contents, they often fail to take actions in a timely fashion [11].

Online social networking sites have developed several mechanisms in order to screen offensive contents in texts. The majority of popular social media use simple lexicon-based approach to filter offensive contents. In order to detect and remove "offensive" language in texts in social media, Z. Xu et al., [12] proposed the Lexical Syntactical Feature (LSF) approach. Using lexical and syntactical features of each sentence derive an offensive value for each sentence. Total three types of features were developed to identify the level of offensiveness, style features, structural features, and content-specific features. Finally examined the classification rates for different feature sets using Naïve Bayes and SVM classifiers. Nahar et al., 2013, [13] proposed an effective approach to detect cyberbullying from social media. And they also presented a graph model to extract cyberbullying network. This has led to identifying the most active predators and victims through a ranking algorithm. Their proposed graph model could be used to recognize the level of cyberbullying victimization for decision making in further studies [14].

For detecting cyberbullying among YouTube comments, Dinakar et al., [15] used a variety of binary and multiclass classifier on a manually labelled dataset. Moreover they applied common sense knowledge for detecting cyberbullying. Using common sense can help provide information about people's goals and emotions and object's properties and relations that can help disambiguate and contextualize language. They also used two types of features: 1) general features that contain a term frequency-inverse document frequency (TF-IDF) weighted uni-grams, the Ortony lexicon of words denoting negative connotation, a list of profane words and frequently occurring part-of-speech (POS) bigram tags and 2) label specific features. Their study indicated that binary classifier can outperform the recognition textual cyberbullying in comparison to multiclass classifiers. Their results illustrate using such features into account will be more useful and can lead to better modelling of the problem. The limitations of their study are that they did not consider the pragmatics of dialogue and conversation and the social networking graph. Recently, Maynard, Bontcheva and Rout (2012) [16] conducted research to detect negative opinions in social media using the rule-based approach. Their research has identified negative opinions, which contain sentiment sentences, with 86% precision and 71% recall, and also identified sentences where the accuracy of the polarity (positive or negative) was 66%.

III. Theoretical Backgound

In the following paragraphs, a brief introduction to core methodologies applied in the current paper is given. More specifically, in order of use by our approach, we provide a short description about SVM, which are applied in order to weight the significance of each feature, SVD, a technology that we use in order to reduce the feature space and DTW, an algorithm that measures the similarity of two signals, which is found to be superior to standard Euclidean distance.

A. Support Vector Machines (SVM)

Support vector machines derived from statistical learning theory by Vapnik and Chervonenkis [17]. Initially SVM popularized in the Neural Information Processing (NIPS) community, now is an important and active field of all Machine Learning research. Basically SVM is a two-class classifier. Given a set of training examples, each marked as belonging to one of two categories, an SVM training algorithm builds a model that assigns new examples into one category or the other, making it a non-probabilistic a binary linear or non-linear classifier. An SVM model is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible [18], [19].

B. Singular Value Decomposition (SVD)

An indexing and retrieval method is Latent semantic indexing (LSI) that uses a mathematical technique called singular value decomposition (SVD) to identify patterns in the relationships between the terms and concepts contained in an unstructured collection of text. LSI is based on the principle that words that are used in the same contexts tend to have similar meanings. The goal of LSI is the extraction of the "meaning" of words by using their co-occurrences with other words that appear in the documents of a corpora. A key feature of LSI is its ability to extract the conceptual content of a body of text by establishing associations between those terms that occur in similar contexts. LSI begins by constructing a term-document

matrix A to identify the occurrences of the m unique terms within a collection of n documents. In a term-document matrix, each term is represented by a row, and each document is represented by a column, with each matrix cell $a_{i,j}$. The singular value decomposition (SVD) is a factorization of a real or complex matrix, with many useful applications in signal processing and statistics [20].

C. Dynamic Time Warping (DTW)

Dynamic Time Warping algorithm is a very popular type of distance which was first introduced to the database community in [21]. DTW calculates an optimal warping path between two series of data points (e.g., time series). A time series is a collection of observations made sequentially in time. Lots of useful information can be obtained by measuring time series data over times. Finding out the similarity between two time series is the heart of many time series data mining applications. DTW between two time series does not require the two series to be of the same length, and it allows for time shifting between the two time series by repeating elements. It matches two time series together by allowing them to stretch, without rearranging the sequence of the elements. Particularity of DTW is that it compares two time series together by allowing a given point from one time series to be matched with one or several points from the other.

IV. Methodology

In this section, the main steps of the proposed mining approach will be described.

A. Data description

All data discussed in the present paper were downloaded from Perverted-Justice (also known as PJ) [22], an American organization, which investigates, identifies, and publicizes the conduct of adults who solicit online sexual conversations with adults posing as minors. The site consists of volunteers who carry out sting operations by posing as minors (the age range chosen for the decoys is 10-15) on chat sites and waiting for adults to approach them. After obtaining identifying information of adults posing as minors, who may offer their telephone numbers and other details so that meetings can be arranged, the organization passes the information on to law enforcement. A possible question about this dataset is to what extent the findings may be an artefact of all these "conversations" being setups, possibly based on a script that the organisation "PervertedJustice" gives to all its volunteers who pose as minors in order to attract and convict cybergrooming individuals. According to the Perverted Justice organization, the complete unedited chat logs, which usually contain sexually explicit content and obscenities (and sometimes annotated with comments from the Perverted-Justice volunteer) are posted to the website only after the person's legal case has been resolved. The current follow-up process consists of notifying a community of the offender's status once a person has been arrested and convicted. In the present method, we examine the questions set of each predator. The goal of the present work is to capture the tone of predator's defensive or aggressive questions in order to identify patterns of predator's behavior that can be generalized in a real -life conversation.

In 2010 Kontostathis et al., [23] used approximately 33 chat logs for the identification of potential "sexual predators" in online conversations which transcript from Perverted Justice and developed a software system called "Chat Coder", which was designed to make a decision for which lines in chat logs contained offensive language. These transcripts are labeled with personal information about the predators. All of the predators who participated in these chats have been convicted. In our experiments, we used exactly the same online dialogs setting and labeling of the data, as our dataset. However, we present a completely different approach, moreover we don't follow exactly the same evaluation procedure in order to achieve compatibility of comparison with previously reported result. Therefore, we cannot directly compare our results with those of previous studies of cyberbullying. This collection of posts provides an insight on the kind of questions a predator asks in order to lure its victim and the type of answers returned. Despite the fact that other information is also available in this dataset (such as first name, last name, online stated name, age, gender, race, city, state, repeated offender, admit guilt, etc.) we are particularly interested in analyzing the predator's post, since for a real-life application, this source of information is always available and more informative. In all questions asked by the predator, a numeric class label was assigned (named as *category*), manually annotated by two trained students of the Department of Media and Communication Studies, Ursinus College. The label contains values from the set {0,200,600,900}. Zero is assigned to posts where no cyberbullying activity or intention is identified. Questions which contain personal information are described as category 200. In this category, indicators such as personal information (age, hobbies) is included. Category 600 is assigned as grooming and characterizes the posts which contain words with sexual meaning (e.g., sex, orgasm, kiss, naked, etc.). Any attempt of the predator to physically approach the victim is labeled as category 900 and includes terms that are associated with such a behavior, such as approach verbs (e.g., come, meet), families nouns (e.g., mom, divorce), an isolation an adjective (alone, lonely), etc. For example, the use of a family noun such as "divorce" could represent a predator's attempt to gain insight on the physical location of the victim's family. Therefore, this word could indicate an attempt from the predator to physically approach the victim, in order to isolate it from its support network of family, friends, etc. [7].

B. Preprocessing

All predator posts from the set of XML transcripts were parsed and transformed into a vector. While in most text mining approaches there are some standard preprocessing steps such as stop-words removal, tokenization, stemming and Part-Of-Speech tagging, analysis of the transcripts revealed that there are certain acronyms and text shorthand used in web chat situations that need to be taken into consideration and would be removed if one applied the aforesaid standard preprocessing procedure. Some characteristic examples of such cases include the use of "121", which is a well-known shorthand for the phrase "one to

one", "182", which is used to represent "I hate you", "ADIDAS" for "All Day I Dream About Sex" and many others. A detailed study of this phenomenon can be found in [24]. For a full list of acronyms, visit: http://www.netlingo.com.

Therefore, the only preprocessing steps we applied was a tokenization based on the space character, a stop-word removal and a case transformation.

C. Data representation

Upon preprocessing of posts, two main models were examined for the vector representation. The former was the simple bag-of-words approach, in which all tokens are considered as input feature while the latter model was chosen to be the n-gram character representation, with n equals to 3. There are certain researches that criticize the use of n-grams of characters such as Chen et al., study [11] and according to our experiments, the latter model actually worsened the performance. The feature representation of text documents plays a critical role in many applications of data-mining. The sparse Bag-of-Words representation is arguably one of the most popular and effective approaches. Each document is represented by a high dimensional sparse vector, where each dimension corresponds to either the term frequency of a unique token or the number of term occurrences or a simple binary representation of the presence or absence of the term. A natural extension is *tf-idf*, where the term frequency counts are discounted by the inverse document-frequencies. For the task at hand, all four variations were examined, namely *tf-idf*, term frequency, term occurrences and binary term occurrences.

More specifically, the set of predator's questions is ordered by date and time in increasing order. Each question is modelled as a vector using the tf-idf weighting scheme. The size of the resulted matrix equals n*(m+1), where n is the number of questions of each predator and m is the number of terms obtained upon the text-processing phase. The +1 is for the class label, which is the numeric score of the level of assault, set by the annotators according to the rules as explained previously at the data description section. Therefore, each set of questions asked by a predator was modeled as a time series ts_j, with j ranging from 1 to 33 (i.e. the number of transcribed logs from the aforementioned dataset) and each ts_j is represented as an $n \times m$ matrix. It is evident that using the specific data representation format, m should be quite large.

D. Feature Weighting

In a plethora of supervised learning problems, data sparsity is likely to occur, caused by the large number of attributes considered. Particularly in applications of text mining, bio-informatics and image processing, this phenomenon may deteriorate classification performance. Since the proposed application is strongly related to text mining, and since the data representation follows the bag-of-words model, it is evident that feature selection can address that issue from a pragmatic view, of improving the prediction outcome and make computations more feasible [25]. In the current work, the selection process is accomplished via a linear SVM weighting methodology. Data instances within a time series ts_j are described as vectors $x_i = (x_{i1}, x_{i2}, \dots, x_{im})$, where m represents the dimensionality of the feature space of ts_j.. The SVM predicts the class of an instance x by the following form:

$$prediction(x) = sgn[b + w^T x], w = \sum_i \alpha_i x_i.$$

Vector $w = (w_1, w_2, \dots, w_m)$ is the projection of the hyper-plane that separates instances according to their class (for a binary classification problem) [25]. The main idea behind the selection process is that one may consider an attribute important if it portrays a substantial impact to the width of the margin of the resulting hyper-plane. Such a margin is inversely proportional to$\|w\|$, i.e. the length of w. taking the fact that $w = \sum_i \alpha_i x_i$into account, one could assume that the influence of a feature $k, with\ k \in \{1,2, \dots, m\}$ on the width of a margin could be estimated by considering the absolute values of partial derivatives of $\|w\|^2$ with respect to x_{ik}. For the linear kernel, this proposition can be written as:

$$\sum_i |\partial \|w^2\| / \partial x_{ik}| = \delta |w_k|,$$

where the sum is on the support vectors and δ is a constant, independent of k. Hence, the features with the higher $|w_k|$are more influential.

E. Dimensionality reduction

The alternative of the SVM feature selection approach we followed in order to alleviate the issue of high-dimensionality of data, was that of SVD [27]. The main idea behind is that SVD can reveal "concepts" i.e. latent relationships between features in an ordered manner, so that one could choose the degree of dimensionality reduction. For our case, we retained enough singular values to make up 90% of the energy in matrix S (as explained in Section III). That is, the sum of the squares of the retained singular values should be at least 90% of the sum of the squares of all singular values.

F. Time Series forecasting

In some past works, it was quite often that whenever feature selection was performed using linear SVM, the same algorithm also served as the classifier. Nevertheless, there are also previous researches that distinguish the two processes and use another algorithm for classification. In our setting, linear as well as non-linear SVM were measured and the performance was benchmarked against that of a Multi-Layer Perceptron (MLP) neural net. In the following section, analytic results tabulate this difference. For the forecasting task, a window size of K previous posts was chosen and the horizon was set to one (i.e. the following post) [28], [29], [30].

The MLP architecture was chosen to contain one hidden layer perceptron with the following data processing function:

$$Y_t^* = \Phi_{output}\left(\sum_{j=1}^{h} u_j \, \Phi_{hidden}\left(\sum_{l=1}^{K}\sum_{i=1}^{m} w_{m(l-1)+i} X_{it-l} + w_0\right) + u_0\right)$$

where Y_t^* is the prediction of the t-th post, X_i is the data for current post i, K is the window size, m is the dimensionality of the input vector, w_i are the MLP hidden layer coefficients, υ_j represent the output layer coefficients, n is the number of inputs, h is the number is neurons within the hidden layer and φ(x) are the activation functions. As regards to the output layer,

$$\Phi_{output}(x) = \left(\frac{1}{1+e^{-x}} - 0.5\right).$$

For the hidden layer, the traditional sigmoid function is used, i.e.:

$$\phi_{hidden}(x) = \frac{1}{1+e^{-x}}.$$

In order to estimate the perceptron's weights, a genetic algorithm was applied, following the approach of Korning PG. study [31].

G. Correlation of SVD dimension and Class distribution

From the initial steps of preprocessing the available data, we observed that the class label was following a frequent pattern behavior. More precisely, predators initiated conversation using generic questions and some personal information (category=0 or category = 200) and afterwards, more aggressive tone was used, most often belonging to the category of 600. Finally, questions that belonged to the last category (900) were asked. Sometimes, between signals of 600 or 900, low intensity questions interfered. This generic observation, along with the fact that categories 200, 600 and 900 were closely associated with specific type of language use, led us to search for possible patterns within linguistic tokens. As expected, the distribution of specific tokens within a transcript was following the same course with their associated class distribution. For example, the plot of word "meet" within a time series was peaking at the same period of category (i.e. a value of 900) and this is of course natural, since human annotators have labeled as 900 all the questions where the predator asks the victim to meet each other.

Similar attitude was appearing for a plethora of terms, from the various class labels. Our main goal was to create a unified signal that would capture the behavior of individual linguistic information as a whole, expecting that such a signal could be highly correlated with the class signal. Indeed, this goal was achieved when applying the SVD method on each time series. As already known, SVD is ideal for dimensionality reduction and noise removal. Apart from that, SVD is able to discover "concepts" within features, meaning that the ranked singular values symbolize the strength of such concepts.

The first SVD dimension that results upon transforming the n*m matrix represents the level of interrelations between the rows (i.e. the questions of the predator) and the columns (i.e. the terms), as set by trivial metrics such as the count of terms and the length of each vector (i.e. how many terms where found in a single question). Therefore, this dimension is of no use for the task at hand. The second, the third and so forth dimensions capture latent semantic relations in a decreasing order of significance. Experimenting with various SVD dimensions, an interesting pattern appeared. More specifically, the second dimension portrayed similar behavior with the class label series. As we shall see in the following experimental section, by considering the second singular value and taking the second row of matrix V^T (denoted as SVD_2) from the SVD procedure, the plot of this vector revealed high degree of similarity with the plot of the class label vector.

Since these two vectors had similar but not identical curve (e.g. there could be a small shift towards time of difference in magnitude) similarity was measured using DTW which is far more robust and accurate than Euclidean distance in time series. This interesting finding allows different types of analysis tasks. For example, one may use unlabeled data, with no class information, and transform them using the 2nd SVD dimension and use this plot for automatic labelling of the unseen data. Moreover, one could compare the SVD time series of an unseen instance with the labeled time series of the training data and find potential repeated offends. One could also exploit the SVD plot in order to predict the future class using the previous values of the plot.

In Fig. 2, an illustration of the similarity of the two signals, namely category (i.e. the class label) and the second SVD dimension. The former is positioned on the vertical axis of the figure while the latter is situated on the horizontal. In the main part of the illustration, we can see the warping matrix, which is used in order to align the two sequences. Finally, the solid white line represents the optimal warping path, which is characteristic indicator of the similarity of category and SVD_2. As we could observe, the two signals expose a great degree of similarity, denoted by the optimal warping path which is quite close to the diagonal of the warping matrix.

A detailed analysis of the correlation monitoring of these sequences from all of the time series is presented in the experimental results section that follows.

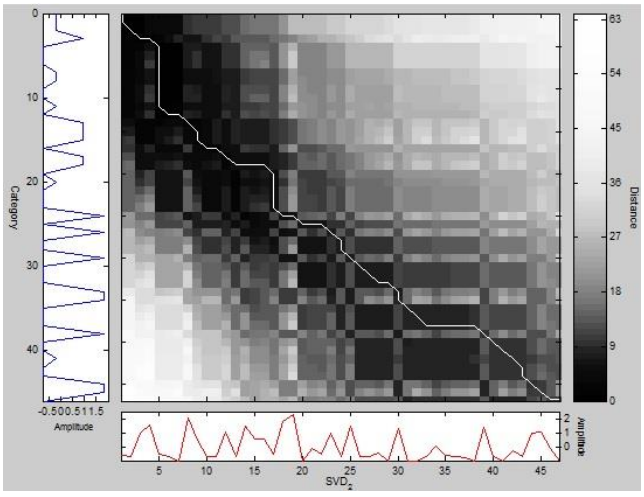

Fig. 2. The most representative example from the time series data (predator: *ArmySgt1961*) where category and SVD2 signals portray high similarity using the DTW algorithm.

V. Experimental Results

In order to validate the outcome of the proposed methodology, two different sets of experiments were carried out. First, the forecasting performance of the category label was measured for each predator, utilizing different feature sets, window sizes and prediction algorithms. The second phase evaluated the correlation of the class sequence with the sequence of SVD_2.

A. Forecasting Performance

Several groups of experiments on each of the 31 transcripts were performed. Prior to the explanation of the experiments and results, the description of the appropriate metrics for comparison between solutions is presented.

1) Evaluation Crtiteria

Two of the most common metrics have been considered, namely the Root Mean Square Error and the Mean Absolute Percent Error [26].

The Root Mean Square Forecasting Error (RMSFE) of an algorithm *I* is calculated by the equation:

$$\text{RMSFE}_{\,i=}\sqrt{\frac{\sum_{j=1}^{n}\left|P_{(ij)}-T_j\right|^2}{n}}$$

where $P_{(ij)}$ is the value that algorithm I forecasted for the sample j (from a set of examples) and T_j is the value of the 'target value' for the j^{-th} example. For an ideal forecast, $P_{(ij)}= T_j$ and $RMSFE_i=0$. So, the error indicator varies from 0 to infinity, with 0 to correspond to the ideal prediction.

The Mean Absolute Percent Error (MAPE) of an algorithm *I* is practically a measure that corrects the 'canceling out' effects of positive and negative errors and also keeps into account the different scales at which this measure can be computed. The MAPE is given by:

$$MAPE_i=\frac{100}{n}\sum_{j=1}^{n}\frac{\left|P_{(ij)}-T_j\right|}{T_j}$$

2) Forecasting methods

Two different classifiers were tested in order to achieve the best prediction outcome. SVMs using linear and polynomial kernels, and MLP Neural Networks. For the SVM and MLP parameters, the optimal parameters were found using a Genetic Algorithm implementation of the RapidMiner® data mining suite. As regards to the dataset, we used the standard bag-of-words approach, the SVM feature weighting keeping from 25%-50% of the original attributes and the SVD transformation using five dimensions. For each of the above feature schemes, four different vector representations were used, such as *tf-idf*, term frequency (tf), term occurrences (to) and binary term occurrence (bto). Finally, two different window sizes were selected (two and three) and horizon was set to the next predator's post category.

The following Table illustrates the average RMSFE of MLP Neural Nets and SVM for window size of two.

TABLE I. **RMSFE** for MLP neural networks and SVM, for all representation formats, using a window of **two**.

	Bag-of-words		**Weights**		**SVD**	
	MLP_NN	SVM	MLP_NN	SVM	MLP_NN	SVM
tf-idf	**0,171**	**0,226**	**0,147**	**0,147**	**0,143**	**0,177**
tf	**0,156**	**0,205**	**0,094**	**0,116**	**0,106**	**0,117**
to	**0,194**	**0,232**	**0,15**	**0,194**	**0,14**	**0,141**
bto	**0,267**	**0,245**	**0,158**	**0,158**	**0,2**	**0,206**

As we can observe, the best results are obtained using MLP neural networks for the feature set which is represented by term frequencies and reduced using SVM feature selection and keeping the 25% of the features. SVM as a predictor is also portraying satisfactory forecasting performance for the same feature set but is significantly reduced when using the bag-of-words approach. The SVD method did better than the initial feature set using both neural nets and SVM but could not outperform the feature selection representation. Similar outcomes are observed when using MAPE as criterion (Table II). Again, MLP neural networks are better than SVM in all representations. Feature selection using SVM and term frequency again portray the best results.

TABLE II. **MAPE** for MLP neural networks and SVM, for all representation formats, using a window of **two**.

	Bag-of-words		**Weights**		**SVD**	
	MLP_NN	SVM	MLP_NN	SVM	MLP_NN	SVM
tf-idf	**34,10%**	**35,40%**	**27,88%**	**29,30%**	**28,50%**	**35,40%**
tf	**30,23%**	**36,45%**	**21,13%**	**25,44%**	**23,10%**	**23,40%**
to	**38,70%**	**36,00%**	**40,00%**	**38,90%**	**34,00%**	**35,00%**
bto	**49,00%**	**40,30%**	**41,00%**	**38,00%**	**43,00%**	**47,00%**

TABLE III**. RMSFE** for MLP neural networks and SVM, for all representation formats, using a window of **three**.

	Bag-of-words		**Weights**		**SVD**	
	MLP_NN	SVM	MLP_NN	SVM	MLP_NN	SVM
tf-idf	**0,154**	**0,175**	**0,045**	**0,050**	**0,100**	**0,097**
tf	**0,152**	**0,157**	**0,041**	**0,044**	**0,085**	**0,110**
to	**0,187**	**0,184**	**0,101**	**0,080**	**0,103**	**0,108**
bto	0,168	0,155	0,114	0,106	0,158	0,144

TABLE IV. **MAPE** for MLP neural networks and SVM, for all representation formats, using a window of **three.**

	Bag-of-words		**Weights**		**SVD**	
	MLP_NN	SVM	MLP_NN	SVM	MLP_NN	SVM
tf-idf	**37,00%**	**34,00%**	**29,40%**	**30,10%**	**26,40%**	**36,40%**
tf	**33,30%**	**42,12%**	**18,80%**	**19,30%**	**21,10%**	**28,00%**
to	**38,70%**	**39,40%**	**30,00%**	**30,05%**	**33,00%**	**36,40%**
bto	**49,80%**	**47,10%**	**31,60%**	**30,33%**	**48,00%**	**48,30%**

In Table III and Table IV, the same metrics and algorithms are applied, however for a longer window size, i.e. three. This time, the bag-of-words is depicting worse results than when using a window of size two. This is actually expected since data dimensionality and sparsity pose significant challenges to both classifiers. Neural networks are again better than SVM for all representation schemes, however, this difference is now smaller than before. More specifically, the difference ranges from 2% to 3.5%. Finally, Table V decomposes performance for the best representation (i.e. Feature Selection using SVM and 25% of the original attribute set, using term frequency) for each predator, using the two classifiers. The color of each column ranges from dark red to dark green, with the worst results denoted by the former and the best by the latter. We have to mention that the perpetrators' pseudonyms are not considered personally identifiable information, as they have already used in other published works [9].

TABLE V. ANALYTICAL PERFORMANCE EVALUATION (RMSFE, MAPE) PER PREDATOR, USING: WINDOW=3, FEATURE SELECTION BY SVM AND TF

	Algorithm			
	MLP_NN		**SVM**	
Name	**RMSFE**	**MAPE%**	**RMSFE**	**MAPE%**
ArmySgt1961	0.040	18.7	0.039	18.7
arthinice	0.042	19	0.042	18.4
asian_kreationz	0.039	18	0.046	19.5
aticloose	0.034	18	0.039	19
corazon23456	0.043	18.2	0.045	19.5
crazytrini85	0.035	17.8	0.042	19.3
flxnonya	0.043	18	0.046	19.2
fotophix	0.043	18.8	0.041	18.8
ghost27_73	0.035	19.1	0.046	18.3
hiexcitement	0.034	18.4	0.043	18.9
i_8u_raw	0.042	18.3	0.043	19.4
icepirate53	0.040	18.8	0.039	19.3
italianlover37	0.040	19.2	0.044	18.7
jleno9	0.044	17.8	0.047	19.1
jon_raven2000	0.044	19.2	0.039	18.7
lee_greer74	0.036	18.5	0.044	18.4
manofdarkneedsl951	0.034	18.4	0.043	19.2
marc_00_48089	0.039	18	0.045	19
needinit1983	0.041	18	0.042	18.9
sebastian_calif	0.042	18.2	0.043	19.2
sjklanke	0.041	19	0.046	19
sphinx_56_02	0.044	19	0.046	18.8
spongebob_giantdick	0.035	17.7	0.042	19.4
stylelisticgrooves	0.036	17.7	0.043	19.1
sugardavis	0.036	18.6	0.044	18.2
sweet_jason002	0.039	19.1	0.040	19
texassailor04	0.041	18.9	0.047	19.1
the_third_storm	0.041	18.8	0.039	19.2
thedude420xxx	0.037	18.3	0.040	18.9
tunnels12000	0.040	18.2	0.042	18.1
user194547	0.038	18.1	0.039	18.2

B. SVD_2-Class correlation

This section presents the experimental outcome of the comparison of the two sequences, i.e. the class category against the SVD_2 dimension of each time series. Each signal was

normalized using a Z-transformation. Additionally, the Euclidean distance was used in order to depict that DTW is a more accurate similarity estimator when dealing with time series. As Table VI tabulates, the DTW similarity is kept at low levels throughout the whole dataset, denoting that the category class could be simulated by a more general term distribution representation such as SVD_2. Again, the color coding helps to identify global minima and maxima per column. This discovery could be used in a variety of ways, such as automatic annotation of chat logs, pattern matching of new, unseen data against existed and proved cases, in order to identify repeated offenders, etc.

VI. Concluding Remarks

The present article presents our study on pattern discovery and evaluation of the strategies used by online sexual predators in their efforts to develop relationships with minors using the Internet. This case is a particular area of a more generic phenomenon, namely Cyberbullying and has been of major significance throughout recent years due to the vast spread of social media networks and mobile devices.

For the purposes of our research, 31 real world transcriptions have been used as source data, obtained from a well-known American organization (i.e. Perverted Justice), which inspects, identifies, and publicizes the conduct of adults who solicit online sexual conversations with adults posing as minors.

Based on the questions asked by each predator and a manual assignment of a category label, denoting the type of attack, analysis was performed in order to model the predator's tactics.

Unlike previous works, the main contribution of this article is that it confront the task as a time series modeling approach, in which previous states portray important information on the knowledge of a future state. Furthermore, another differentiation from other approaches is that we consider an initial bag-of-words document representation method, containing idioms, slang, emoticons and abbreviations. Other approaches use more structured forms for feature representation such as inclusion of swear words, domain-specific vocabulary, etc.

However, this choice for feature representation resulted in sparse data instances. In order to alleviate this problem, two different strategies were followed: the former utilized feature weighting and selection by a linear SVM algorithm while the latter exploited linear algebra dimensionality reduction methods such as SVD. Experimental results from the two strategies using advanced forecasting techniques have portrayed very satisfactory performance for some cases.

A third novelty of the paper is the identification of similar behavior between the sequences of the class label and the second SVD dimension of the input vector (i.e. the questions posed by the predator). Using DTW, this pattern was verified in for a large proportion of the dataset, indicating a straightforward alignment between the type of attack and the language used throughout the course of a dialogue. This alignment may be used for discovery of repeated offends by a same user, or for clustering of similar cases or even for automatic annotation of unlabeled data. The resemblance of the 2nd SVD dimension with the class label could also be exploited in generalizing the classification method for cyber monitoring real-life conversations.

TABLE VI. Results on the similarity of the Class label (i.e. Category) and the SVD_2 dimension of the feature set.

	Similarity	
Name	**DTW**	**EUCLID**
ArmySgt1961	2.98	6.38
arthinice	2.74	5.78
asian_kreationz	1.81	3.80
aticloose	2.31	4.94
corazon23456	1.93	4.19
crazytrini85	2.07	4.45
flxnonya	2.51	5.37
fotophix	1.98	4.30
ghost27_73	2.64	5.68
hiexcitement	2.72	5.93
i_8u_raw	1.81	3.98
icepirate53	2.17	4.75
italianlover37	2.93	6.36
jleno9	2.83	6.06
jon_raven2000	2.5	5.38
lee_greer74	2.61	5.69
manofdarkneedsl951	1.81	3.96
marc_00_48089	2.08	4.49
needinit1983	2.96	6.28
sebastian_calif	2.91	6.40
sjklanke	2.93	6.18
sphinx_56_02	2.7	5.67
spongebob_giantdick	2.7	5.67
stylelisticgrooves	2.61	5.74
sugardavis	2.17	4.77
sweet_jason002	2.49	5.30
texassailor04	2.97	6.53
the_third_storm	2.04	4.32
thedude420xxx	3.04	6.57
tunnels12000	2.96	6.33
user194547	3.03	6.51

Even if the examined corpus contains identified predator's posts, each post is of a varying bullying significance, ranging from a typical, non-suspicious question to questions with sexual or harassing content. Therefore, by observing the SVD plot, one could identify regular dialogues (i.e. when the curve is almost flat line) or posts with potential bullying terms, where in such cases the SVD plot would present peaks or troughs.

As regards to future directions, we are planning to further investigate this domain by also considering the user response as part of the input vector and try more advanced feature representation techniques such as topic modeling utilizing the Latent Dirichlet allocation methodology.

REFERENCES

[1] Cyberbullying, The National Crime Prevention", [online] http://www.ncpc.org/cyberbullying.

[2] Willard, N. E. (2007). Cyberbullying and Cyberthreats: Responding to the Challenge of Online Social Aggression, Threats, and Distress. Champaign, IL: Research. Print.

[3] E. Cambria, H. Wang, B. White. Guest Editorial: Big Social Data Analysis. Knowledge-Based Systems 69, pp. 1-2 (2014).

[4] Inches, G., Crestani, F.: Overview of the international sexual predator identification competition at pan-2012. In: Forner, P., Karlgren, J., Womser-Hacker, C. (eds.) CLEF 2012 Evaluation Labs and Workshop - Working Notes Papers. Rome, Italy (2012).

[5] Yin, Z. Xue, L. Hong, B. D. Davison, A. Kontostathis, and L. Edwards. (2009). Detection of Harassment on Web 2.0 in CAW 2.0 '09: Proceedings of the 1st Content Analysis in Web 2.0 Workshop, Madrid, Spain.

[6] Kontostathis, A., Reynolds, K., Garron, A., & Edwards, L. (2013). Detecting cyberbullying: query terms and techniques. In Proceedings of the 5th annual acm web science conference (pp. 195–204).

[7] Kontostathis, A., West, W., Garron, A., Reynolds, K., Edwards, L.: Identifying Predators Using Chatcoder 2.0 - notebook for pan at clef 2012.

[8] R. Lau, Y. Xia, Y. Ye. A probabilistic generative model for mining cybercriminal networks from online social media. IEEE Computational Intelligence Magazine 9(1), pp. 31-43 (2014).

[9] Dadvar, M., F. de Jong, Ordelman, R. and Trieschnigg, D. 2012. Improved Cyberbullying Detection Using Gender Information, In Proceedings of the 12th -Dutch-Belgian Information Retrieval Workshop(DIR2012) (Ghent, Belgium 2012).

[10] McGhee, I., Bayzick, J., Kontostathis, A., Edwards, L., McBride, A., Jakubowski, E.: Learning to Identify Internet Sexual Predation. International Journal of Electronic Commerce 15(3), 103–122 (Apr 2011).

[11] Chen, Y., Zhou, Y., Zhu, S.,& Xu, H. (2012) : 'Detecting Offensive Language in Social Media to Protect Adolescent Online Safety', in Privacy, Security, Risk and Trust (PASSAT), 2012 Internationa Conference on and 2012 International Confernece on Social Computing (SocialCom), pp. 71-80.

[12] Z. Xu and S. Zhu. Filtering offensive language in online communities using grammatical relations. In Proceedings of the Seventh Annual Collaboration, Electronic Messaging, Anti-Abuse and Spam Conference, 2010.

[13] V. Nahar, X. Li, C. Pang, "An Effective Approach for Cyberbullying Detection". Journal of Communications in Information Science and Management Engineering .2013.

[14] Samaneh Nadali, Masrah Azrifah Azmi Murad, Nurfadhlina Mohamad Sharef, Aida Mustapha,Somayeh Shojaee, "A Review of Cyberbullying Detection : An Overview" 13th International Conference on Intelligent Systems Design and Applications (ISDA),2013.

[15] K. Dinakar, R. Reichart, and H. Lieberman, "Modeling the Detection of Textual Cyberbullying," International Conference on Weblog and Social Media - Social Mobile Web Workshop, Barcelona, Spain 2011, 2011.

[16] D. Maynard, K. Bontcheva, and D. Rout. Challenges in developing opinion mining tools for social media. In Proceedings of the @NLP can u tag #usergeneratedcontent?! workshop, LREC, pages 15–22,2012.

[17] Vapnik, V. (1995). The natural of statistical Learning Theory. Springer, New York.

[18] A. Rakotomamonjy,"Variable selection using SVM-based criteria,"J.Mach. Learn. Res., vol. 3, pp. 1357–1370, 2003.

[19] Joachims, T. Text categorization with support vector machines. Technical report, LS VIII Number 23, University of Dortmund, 1997.ftp://ftpai.informatik.uni/dortmund.de/pub/Reports/report23.ps.Z.

[20] Kalman,D.1996.A singularly valuable decomposition:the SVD of a matrix.College Math.J.27:2-23.

[21] Keogh, E. & M. Pazzani. Derivative Dynamic Time Warping. In Proc. of the First Intl. SIAM Intl. Conf. on Data Mining, Chicago, Illinois, 2001.

[22] [online], http://www. Perverted-Justice.com.,2008.

[23] A. Kontostathis, L. Edwards, and A. Leatherman, "ChatCoder: Toward the Tracking and Categorization of Internet Predators," In Proceedings of Text Mining Workshop 2009 held in conjunction with the Ninth SIAM International Conference on Data Mining (SDM 2009).

[24] K. Dinakar, B.Jones, C.Havasi, H.Lieberman, and R.Picard, "Common sense reasoning for detection, prevention, and mitigation of cyberbullying," ACM Trans. Interact. Intell. Syst., vol. 2, no. 3, pp. 29–31, 2012.

[25] Brank, Janez, Marko Grobelnik, Natasa Milic-Frayling, and Dunja Mladenic. (2002). "Feature Selection Using Linear Support Vector Machines." Technical report, Microsoft Research.

[26] Chang Y-w, Lin C-j: Feature ranking using linear SVM. J Machine Learning Res 2008,3:53-64.

[27] PLATT, J. C. 2000. Probabilistic outputs for support vector machines and comparison to regularized like-lihood methods. In Advances in Large Margin Classifiers, A. Smola, P. Bartlett, B. Scholkopf, and D. Schuurmans, Eds. MIT Press, Cambridge, MA.

[28] Raudys, A., & Mockus, J. (1999). Comparison of arma and multilayer perceptron based methods for economic time series. Informatica,10 (2), 231–243. ISSN 0860-4952.

[29] Kim, T., & Adali, T. (2002). Fully complex multi-layer perceptron network for nonlinear signal processing. Journal of VLSI Signal Processing, 32, 29–43.

[30] Weston, S.Mukherjee, O. Chapelle, M.Pontil, T.Poggio,and V. Vapnik. Feature selection for SVMs. In NIPS 13, 2000.

[31] Korning PG., 1994.Training neural networks by means of genetic algorithms working on very long chromosomes. Technical Report. Cmoputer science Department. Aahus C, Denmark.